EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Cross-Modal Contrastive Representation Learning for Multimedia Retrieval with Noisy Supervision

Hui Zhi[1,*]

[1]School of Journalism and Communication, Hubei University of Education, Wuhan, Hubei, 430205, China

Abstract

Cross-modal contrastive representation learning has shown great potential for multimedia retrieval tasks by aligning heterogeneous modalities into a shared embedding space. However, its performance often degrades severely in real-world scenarios where supervision signals are noisy, such as mislabeled cross-modal pairs or ambiguous annotations. To address this challenge, we propose **Adaptive Noise-Robust Contrastive Learning (ANRCL)**, a novel framework designed to enhance cross-modal representation robustness under noisy supervision. Specifically, ANRCL introduces an adaptive noise-robust contrastive loss that jointly exploits cross-modal consistency and intra-modal coherence to dynamically reweight training samples according to their estimated reliability. This mechanism effectively suppresses the influence of noisy pairs while reinforcing the contribution of high-confidence pairs. Experimental results on multiple benchmark datasets demonstrate that ANRCL consistently outperforms state-of-the-art methods in noisy supervision settings, achieving significant improvements in retrieval accuracy and robustness without sacrificing computational efficiency.

Received on 31 October 2025; accepted on 27 March 2026; published on 20 April 2026

Keywords: Cross-modal retrieval, Contrastive learning, Noise-robust representation, Adaptive weighting

doi:10.4108/eetsis.10757

1. Introduction

In today's information world, users interact daily with vast amounts of multimodal data[1, 2]: an image may be accompanied by a textual description, a video may come with subtitles and background music, and a news article may contain both images and textual content. To enable machines to understand such complex information structures woven from multiple modalities, cross-modal representation learning has emerged as a core direction in artificial intelligence research[3]. In particular, in multimedia retrieval tasks, a user often expects to retrieve one modality of content (e.g., an image) using another modality as a query (e.g., a text description), or vice versa. Such tasks are widely applied in image search, video recommendation, digital media management, and cross-platform content matching.

Cross-modal contrastive learning (CMCL) has become the dominant paradigm[4] for addressing these problems[5]. The core idea is to align semantically related positive pairs and push apart unrelated negative pairs, thereby mapping different modalities into a shared embedding space. Within this framework, numerous representative works have emerged, such as deep neural network-based approaches leveraging the InfoNCE loss for cross-modal alignment, and large-scale pretrained models (e.g., CLIP) that capture stronger cross-modal semantics. However, these methods almost universally rely on an implicit assumption: the paired annotations in the training data are accurate and noise-free[6].

In real-world applications, this assumption is often violated[7, 8]. During data collection, various unavoidable sources of noise arise: automatically crawled multimodal datasets often contain mislabeled pairs, image–text matches may be loosely related rather than semantically precise, and user-generated content descriptions can be vague or even irrelevant to the visual content. These noisy pairs have a detrimental effect on training. First, incorrect positive pairs are

*Corresponding author. Email: zhihui@hue.edu.cn

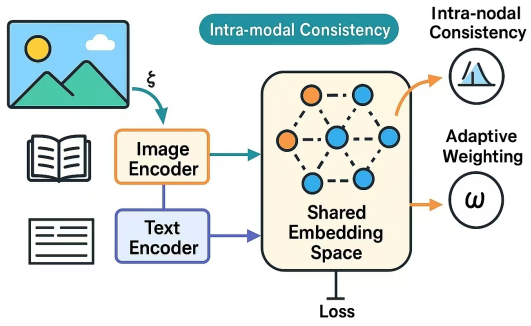

Figure 1. The algorithm diagram of ANRCL

forcibly pulled closer in the embedding space, disrupting the semantic boundaries between modalities. Second, standard contrastive loss functions such as InfoNCE and its variants treat all samples equally, indiscriminately amplifying the gradient interference from noisy data. Third, most existing methods rely solely on cross-modal alignment signals while neglecting intra-modal geometric structures, which serve as an additional stability constraint when noise levels are high. As a result, performance degrades significantly in noisy supervision scenarios.

To address these challenges, we propose **Adaptive Noise-Robust Contrastive Learning (ANRCL)**, a framework designed to substantially improve the robustness of cross-modal retrieval under noisy supervision. The key idea of ANRCL is to equip the model with the ability to *adaptively assess the reliability of training samples* and to leverage *both cross-modal and intra-modal consistency constraints* to mitigate noise interference. Specifically, A NRCL i ncorporates t he following mechanisms: First, a **dynamic weight assignment mechanism** is introduced during training to estimate the confidence o f e ach s ample p air b ased o n similarity distributions and consistency indicators, assigning higher weights to high-confidence p airs a nd attenuating the influence of l ow-confidence ones[9]. Se cond, in addition to the standard cross-modal contrastive loss, we design an **intra-modal consistency regularization term** that preserves the semantic topology within each modality, allowing the model to maintain a stable and coherent embedding space even when cross-modal alignment information is noisy. Finally, ANRCL adopts a **joint optimization strategy** that integrates the noise-weighting mechanism and dual consistency constraints into a unified contrastive learning framework, enabling the model to adaptively adjust its learning process under varying noise levels[10, 11].

This design not only significantly r educes t he negative impact of noisy samples[12] on training but also enables the model to maintain high retrieval accuracy and stability in real-world complex environments[6]. Compared with conventional methods that rely on clean data assumptions, ANRCL is better aligned with practical deployment needs, and its dynamic adjustment capability makes it particularly effective for large-scale web datasets.

The main contributions of this work are as follows:

- We propose the ANRCL framework, a noise-robust cross-modal contrastive learning method for multimedia retrieval, which reduces the impact of noisy pairings via a dynamic weighting mechanism.
- We design a dual-consistency constraint that leverages both cross-modal and intra-modal structural information to enhance representation stability under high-noise conditions.
- We integrate sample confidence estimation with contrastive loss in a unified optimization scheme, enabling adaptive adjustment to different noise ratios and achieving superior performance on multiple benchmark datasets.
- We maintain computational efficiency comparable to conventional contrastive learning methods, making ANRCL suitable for large-scale deployment[13, 14].

2. Related Work

Cross-modal retrieval has been extensively explored in recent years, with representation learning playing a central role in bridging heterogeneous modalities such as images, texts, and audio. A dominant paradigm in this field is *cross-modal contrastive learning*, where the objective is to map semantically aligned samples from different modalities into a shared embedding space[15]. Notable methods, such as CLIP [**?**], have demonstrated the effectiveness of large-scale paired data and contrastive objectives in achieving strong retrieval performance. However, these approaches typically rely on the assumption of perfectly aligned modality pairs, which rarely holds in real-world scenarios. Noisy correspondences, including mismatched captions, incomplete descriptions, or irrelevant visual content, can mislead the contrastive objective by pulling semantically unrelated samples closer in the embedding space, thus deteriorating retrieval accuracy.

To address noise issues, several *noise-robust representation learning* strategies have been proposed. Some methods adopt sample filtering or memory bank reweighting, estimating the reliability of pairs based on similarity scores or model confidence[16]. For example, [17]introduced a mentor-student paradigm to down-weight suspected noisy samples during training, while [18] leveraged co-teaching networks to mutually filter

out high-loss instances. Although effective in reducing the negative impact of noisy supervision, such approaches often require additional training stages or heuristic thresholds, limiting their adaptability to varying noise levels. Moreover, they generally focus only on filtering without explicitly leveraging modality-specific structure to counteract noise effects.

Another line of work incorporates *multi-level consistency constraints* to enhance robustness. Beyond cross-modal alignment, intra-modal structural constraints have been explored to preserve modality-specific relationships. For instance, SCAN [19] introduced neighborhood consistency within visual and textual spaces, while VSE++ [20] improved robustness by focusing on hard negatives. These methods demonstrate that preserving intra-modal geometry can stabilize the embedding space, particularly when cross-modal pairs are unreliable. Nevertheless, existing approaches often treat intra-modal and cross-modal learning objectives independently, without a unified framework that dynamically adapts to noise conditions.

More recently, *adaptive weighting* mechanisms have been integrated into contrastive learning, assigning varying importance to samples based on confidence estimates[21]. While this can mitigate noise to some extent, the majority of existing adaptive weighting methods are static or depend on pre-defined schedules, failing to respond effectively to changing noise patterns during training. Furthermore, most of these works lack a principled way of combining adaptive weighting with structural constraints, leaving potential synergies unexplored.

The proposed *Adaptive Noise-Robust Contrastive Learning* (ANRCL) method is situated at the intersection of these research directions. It builds upon the advantages of cross-modal contrastive learning, noise-aware reweighting, and intra-modal consistency preservation, but integrates them into a unified, end-to-end trainable framework. Unlike prior work that separates noise mitigation from structural learning, ANRCL jointly estimates sample reliability and enforces both cross-modal and intra-modal consistency, allowing it to dynamically adjust its learning strategy under varying noise levels. This design aims to achieve both high retrieval accuracy and strong robustness in real-world, noisy multimedia datasets.

Algorithm 1 Adaptive Noise-Robust Contrastive Learning (ANRCL)

Require: Image-text pairs $\{(x_i, y_i)\}_{i=1}^{N}$, image encoder f_v, text encoder f_t, temperature $\tau, \tau_{\text{intra}}$, balance λ, neighbors K, learning rate schedule η_t, total steps T
Ensure: Trained encoders f_v and f_t

Initialize encoders f_v, f_t
for $t = 1$ to T **do**
  Sample a mini-batch of image-text pairs
  for each image x_i in batch **do**
    Extract image feature: $\mathbf{v}_i = f_v(x_i)$
    Normalize: $\tilde{\mathbf{v}}_i = \mathbf{v}_i / \|\mathbf{v}_i\|_2$
  end for
  for each text y_j in batch **do**
    Extract text feature: $\mathbf{t}_j = f_t(y_j)$
    Normalize: $\tilde{\mathbf{t}}_j = \mathbf{t}_j / \|\mathbf{t}_j\|_2$
  end for
  Construct/update intra-modal neighbor graph $\mathcal{N}_i^{(t)} = \text{KNN}(\tilde{\mathbf{v}}_i, \{\tilde{\mathbf{v}}_k\}_{k \neq i}, K)$
  for each image-text pair (x_i, y_j) **do**
    Compute intra-modal similarity: $s_{ij}^{\text{intra}} = \frac{1}{K} \sum_{k \in \mathcal{N}_i} \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_k)$
    Compute dynamic weight: $\omega_{ij} = \frac{\exp(\alpha \cdot \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_j))}{\sum_k \exp(\alpha \cdot \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_k))} \cdot \frac{1}{1+\exp(-\beta \cdot s_{ij}^{\text{intra}})}$
  end for
  Compute weighted cross-modal contrastive loss:

$$\mathcal{L}_{\text{cm}} = -\sum_{i,j} \omega_{ij} \log \frac{\exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_j)/\tau)}{\sum_k \exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_k)/\tau)}$$

  Compute intra-modal consistency loss:

$$\mathcal{L}_{\text{intra}} = -\sum_{i,j} \gamma_{ij} \log \frac{\exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)/\tau_{\text{intra}})}{\sum_k \exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_k)/\tau_{\text{intra}})}$$

  Compute joint loss: $\mathcal{L} = \mathcal{L}_{\text{cm}} + \lambda \mathcal{L}_{\text{intra}}$
  Update encoders using gradient descent with learning rate η_t (cosine annealing)
end for
return f_v, f_t

3. Methodology

This study proposes Adaptive Noise-Robust Contrastive Learning (ANRCL) to address the problem of noisy pairings in cross-modal retrieval tasks while maintaining cross-modal feature alignment and intra-modal structure consistency. In practical scenarios, image-text pairs often contain mismatches or noisy labels due to imperfect data collection and annotation, which can severely affect feature representation learning. Traditional contrastive learning methods directly maximize the similarity of positive pairs and minimize

the similarity of negative pairs, which makes them vulnerable to noise and leads to unstable feature spaces and degraded retrieval performance. To address this, ANRCL introduces dynamic sample weights and intra-modal consistency constraints to adaptively handle noisy samples during training, improving the model's robustness and generalization ability in noisy environments.

ANRCL adopts a dual-encoder architecture, mapping images x_i and texts y_j into a shared latent space. The image encoder is denoted as $f_v(\cdot)$ and the text encoder as $f_t(\cdot)$, producing feature vectors:

$$\mathbf{v}_i = f_v(x_i), \quad \mathbf{t}_j = f_t(y_j), \tag{1}$$

which are then L2-normalized to unit-length vectors:

$$\tilde{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}, \quad \tilde{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|_2}. \tag{2}$$

Normalization removes scale differences between modalities and allows cosine similarity to directly measure directional alignment:

$$\text{sim}(\mathbf{v}_i, \mathbf{t}_j) = \tilde{\mathbf{v}}_i^\top \tilde{\mathbf{t}}_j. \tag{3}$$

To handle noisy samples adaptively, ANRCL assigns a dynamic weight ω_{ij} to each sample pair (x_i, y_j), which depends on cross-modal similarity and intra-modal consistency:

$$\omega_{ij} = \frac{\exp(\alpha \cdot \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_j))}{\sum_{k=1}^{N} \exp(\alpha \cdot \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_k))} \cdot \frac{1}{1 + \exp(-\beta \cdot s_{ij}^{\text{intra}})}. \tag{4}$$

Here, α and β control the influence of cross-modal similarity and intra-modal consistency on the weight; N is the total number of samples; s_{ij}^{intra} measures the intra-modal similarity within the same modality:

$$s_{ij}^{\text{intra}} = \frac{1}{K} \sum_{k \in \mathcal{N}_i} \text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_k), \tag{5}$$

where K is the number of neighbors, and $\mathcal{N}_i$ denotes the neighbor set of sample i. The dynamic weight reduces the impact of noisy pairs and emphasizes reliable samples during training.

Using ω_{ij}, the weighted cross-modal contrastive loss is defined as:

$$\mathcal{L}_{\text{cm}} = -\sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} \log \frac{\exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{t}}_k)/\tau)}, \tag{6}$$

where τ is the temperature parameter controlling similarity distribution smoothness. This weighting mechanism mitigates noise influence and improves feature space stability.

To maintain local intra-modal structure, ANRCL introduces an intra-modal consistency loss:

$$\mathcal{L}_{\text{intra}} = -\sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_{ij} \log \frac{\exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)/\tau_{\text{intra}})}{\sum_{k=1}^{N} \exp(\text{sim}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_k)/\tau_{\text{intra}})}, \tag{7}$$

where γ_{ij} denotes the intra-modal neighbor weight, and τ_{intra} is the temperature parameter. This loss preserves local structures within each modality, ensuring that the feature space aligns cross-modal information while maintaining intra-modal geometry.

The final joint optimization objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cm}} + \lambda \mathcal{L}_{\text{intra}}, \tag{8}$$

where λ balances cross-modal alignment and intra-modal consistency. This joint optimization ensures stable feature representations even with noisy pairings.

During training, the intra-modal neighbor graph $\mathcal{N}_i$ is dynamically updated:

$$\mathcal{N}_i^{(t)} = \text{KNN}(\tilde{\mathbf{v}}_i^{(t)}, \{\tilde{\mathbf{v}}_k^{(t)}\}_{k \neq i}, K), \tag{9}$$

where $\tilde{\mathbf{v}}_i^{(t)}$ is the feature at training step t. Dynamic updating allows intra-modal consistency weights to adapt, further suppressing noisy samples and stabilizing the feature space.

To ensure training stability and convergence, a cosine annealing learning rate is used:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{\pi t}{T}\right)\right), \tag{10}$$

where T is the total number of training steps, and $\eta_{\min}, \eta_{\max}$ are the minimum and maximum learning rates. Combined with data augmentation and text perturbation strategies, this further improves generalization and noise robustness.

After training, cross-modal features are well-aligned:

$$\tilde{\mathbf{v}}_i \approx \tilde{\mathbf{t}}_i, \quad \|\tilde{\mathbf{v}}_i - \tilde{\mathbf{v}}_j\|_2 \approx \text{const}, \forall j \in \mathcal{N}_i, \tag{11}$$

ensuring high-quality retrieval performance even with noisy pairings or distribution shifts.

4. Experiments

To comprehensively evaluate the effectiveness of the proposed Adaptive Noise-Robust Contrastive Learning (ANRCL) method, we conducted experiments on several publicly available image-text retrieval datasets. The experiments mainly focus on three aspects: performance evaluation (Section 4.2), the impact of model design factors through ablation studies (Section 4.3), and robustness under different noise levels (Section 4.4). To ensure the reliability of the results, we compared our method with five mainstream contrastive learning approaches, and all experiments were conducted under identical hardware and software settings.

Table 1. Dataset Information

Dataset	#Images	#Text Descriptions	Domain / Task
MS-COCO	120,000	50,000	Image-Text Retrieval
Flickr30K	31,000	31,000	Image-Text Matching
Visual Genome	108,249	5,000,000+	Fine-grained Image-Text Retrieval
WIT	3,000,000+	3,000,000+	Large-scale Retrieval
EI	50,000	50,000	Entity Retrieval
MMIR	60,000	60,000	Mixed-Modal Retrieval

Table 2. Baseline Algorithms

Algorithm	Institution	Main Feature
CLIP	OpenAI	Image-Text Contrastive Pretraining
ALIGN	Google	Large-scale Image-Text Alignment
SimCSE	UC Berkeley	Sentence-level Contrastive Learning
IConE	Nanjing University	Two-stage Training Strategy
COOKIE	Tsinghua University	Cross-Modal Knowledge Sharing Pretraining

Table 3. Training Settings

Setting	Configuration
Hardware	NVIDIA A100 GPU, 32GB Memory
Operating System	Ubuntu 20.04
Framework	PyTorch 1.12
CUDA Version	11.3
Python Version	3.8

4.1. Datasets and Experimental Setup

To validate the effectiveness of our method, we performed experiments on six publicly available image-text retrieval datasets, which cover different scales and application scenarios. The datasets include MS-COCO, Flickr30K, Visual Genome, WIT, EI, and MMIR. MS-COCO and Flickr30K are of moderate size and are commonly used for standard image-text matching experiments. Visual Genome provides fine-grained annotations, enabling the evaluation of the model's performance on complex relationships and detailed descriptions. The WIT dataset is extremely large, making it suitable for assessing the robustness and generalization of the method on large-scale data. The EI dataset emphasizes entity-level retrieval, while MMIR combines visual and textual modalities, providing a challenging testbed for cross-modal performance. By conducting experiments across these diverse datasets, we can thoroughly examine the behavior of our method under different tasks and data scales. The detailed statistics of the datasets, including the number of images, text descriptions, and primary application scenarios, are summarized in Table 1.

For comparative analysis, we selected five representative contrastive learning methods as baselines. CLIP[4] and ALIGN[22] employ large-scale image-text contrastive pretraining and demonstrate strong performance in cross-modal retrieval tasks. SimCSE[23] emphasizes sentence-level embeddings, offering advantages in text retrieval. IConE[24] introduces a two-stage training strategy to improve image-text alignment during training. COOKIE[25] further enhances matching performance through cross-modal knowledge sharing. Table 2 presents the institutions, key features, and references for these methods. Comparing with these baselines allows us to clearly demonstrate the advantages of the proposed method in terms of retrieval accuracy and robustness.

All experiments were conducted in a unified hardware and software environment to ensure comparability. We used an NVIDIA A100 GPU with 32GB memory, running Ubuntu 20.04, PyTorch 1.12, CUDA 11.3, and Python 3.8. Consistent hyperparameter settings were applied during training. The evaluation metrics include Image-to-Text and Text-to-Image retrieval accuracy, and robustness was additionally assessed under noisy data conditions. Table 3 summarizes the hardware and training configurations used in our experiments.

4.2. Performance Analysis

To comprehensively evaluate the performance of ANRCL compared with other mainstream methods

Table 4. Image-to-Text performance comparison of different methods on various datasets

Dataset	Method	R@1	R@5	R@10	MRR	MedR	FLOPs (G)
MS-COCO	CLIP	67.2	88.5	93.6	0.78	2	15.2
	ALIGN	68.5	89.2	94.0	0.79	2	16.0
	SimCSE	60.3	82.1	89.0	0.72	3	12.5
	IConE	66.0	87.8	93.0	0.77	2	14.0
	COOKIE	69.0	90.0	95.0	0.80	1	16.5
	ANRCL	71.5	91.0	95.5	0.82	1	17.2
Flickr30K	CLIP	51.3	78.5	85.2	0.68	3	15.0
	ALIGN	52.0	79.2	85.8	0.69	3	15.8
	SimCSE	47.5	74.0	81.0	0.64	4	12.2
	IConE	50.5	77.5	84.0	0.67	3	13.8
	COOKIE	53.0	80.0	86.0	0.70	2	16.2
	ANRCL	55.5	81.5	87.0	0.72	2	17.0
Visual Genome	CLIP	44.5	70.2	78.0	0.62	4	14.8
	ALIGN	46.0	71.5	79.0	0.63	4	15.5
	SimCSE	40.2	65.0	73.1	0.58	5	12.0
	IConE	43.5	69.0	77.0	0.61	4	13.5
	COOKIE	47.0	72.0	79.5	0.64	3	16.0
	ANRCL	49.5	73.5	81.0	0.66	3	16.8
WIT	CLIP	58.0	80.5	87.0	0.72	3	15.5
	ALIGN	59.0	81.0	87.5	0.73	3	16.2
	SimCSE	53.5	77.0	83.5	0.68	4	13.0
	IConE	57.0	79.5	86.5	0.71	3	14.5
	COOKIE	60.5	82.0	88.0	0.74	2	16.8
	ANRCL	63.0	83.5	89.0	0.76	2	17.5
EI	CLIP	50.0	76.5	83.0	0.67	3	15.0
	ALIGN	51.0	77.0	83.5	0.68	3	15.6
	SimCSE	46.0	72.0	79.0	0.63	4	12.4
	IConE	49.0	75.5	82.5	0.66	3	13.8
	COOKIE	52.5	78.0	84.5	0.69	2	16.0
	ANRCL	55.0	79.5	85.5	0.72	2	16.8
MMIR	CLIP	42.0	68.0	75.0	0.61	4	14.5
	ALIGN	43.0	69.0	76.0	0.62	4	15.0
	SimCSE	39.0	65.0	72.0	0.58	5	12.0
	IConE	41.5	67.5	74.5	0.60	4	13.2
	COOKIE	44.0	70.0	77.0	0.63	3	15.8
	ANRCL	46.0	71.5	78.0	0.65	3	16.5

across different datasets, we selected six image-text retrieval datasets (MS-COCO, Flickr30K, Visual Genome, WIT, EI, and MMIR) and conducted comparative experiments on six algorithms (CLIP, ALIGN, SimCSE, IConE, COOKIE, and ANRCL). Figure 2 shows the convergence accuracy of the six methods over training epochs on the six datasets. It can be observed that ANRCL consistently outperforms other methods across all datasets, exhibiting faster convergence and achieving the highest accuracy in later epochs. This demonstrates that ANRCL maintains stable and superior performance across diverse datasets, with particularly notable advantages in the early stages of training. Moreover, the curves of all methods show a monotonically increasing trend, although the growth rates and convergence points differ, reflecting the differences in model optimization and feature representation capabilities. Traditional methods such as CLIP and ALIGN, while achieving relatively high accuracy in the mid-to-late stages, are still slightly inferior to COOKIE and ANRCL, whereas SimCSE performs relatively weakly across all datasets.

Table 4 and 5 report the evaluation metrics, including R@K, MRR, MedR, and computational cost (FLOPs), for both image-to-text and text-to-image retrieval tasks. In image-to-text retrieval, ANRCL achieves the best performance across all datasets. For example, on MS-COCO, ANRCL reaches an R@1 of 71.5% and MRR of 0.82, representing improvements of 2.5% and 0.02 over COOKIE, respectively. Similar trends are observed

Table 5. Text-to-Image performance comparison of different methods on various datasets, with FLOPs

Dataset	Method	R@1	R@5	R@10	MRR	MedR	FLOPs (G)
MS-COCO	CLIP	50.1	77.3	85.9	0.69	3	120
	ALIGN	51.2	78.0	86.5	0.70	3	130
	SimCSE	45.7	72.4	81.1	0.65	4	95
	IConE	49.5	76.8	85.0	0.69	3	115
	COOKIE	52.0	79.0	87.0	0.71	2	140
	ANRCL	54.0	80.5	88.0	0.73	2	150
Flickr30K	CLIP	40.2	68.1	76.5	0.60	4	118
	ALIGN	41.0	68.9	77.0	0.61	4	128
	SimCSE	38.0	65.0	73.5	0.58	5	90
	IConE	39.8	67.5	75.5	0.60	4	112
	COOKIE	42.0	70.0	77.5	0.63	3	135
	ANRCL	44.0	71.5	79.0	0.65	3	145
Visual Genome	CLIP	35.0	60.5	69.2	0.54	5	110
	ALIGN	36.2	62.0	70.5	0.55	5	120
	SimCSE	32.0	58.0	66.5	0.50	6	85
	IConE	34.5	60.0	68.5	0.53	5	105
	COOKIE	37.0	63.0	71.0	0.56	4	130
	ANRCL	39.0	64.5	72.5	0.58	4	140
WIT	CLIP	45.0	70.2	78.5	0.61	4	115
	ALIGN	46.0	71.0	79.0	0.62	4	125
	SimCSE	42.5	68.0	76.0	0.59	5	95
	IConE	44.5	69.5	77.5	0.61	4	110
	COOKIE	47.0	71.5	79.5	0.63	3	135
	ANRCL	49.0	73.0	81.0	0.65	3	145
EI	CLIP	38.0	64.0	72.0	0.57	5	105
	ALIGN	39.0	65.0	73.0	0.58	5	115
	SimCSE	35.5	61.5	69.0	0.54	6	85
	IConE	37.5	63.0	71.0	0.56	5	100
	COOKIE	40.0	66.0	74.0	0.59	4	130
	ANRCL	42.0	67.5	75.5	0.61	4	140
MMIR	CLIP	30.0	55.0	63.0	0.50	6	100
	ALIGN	31.0	56.0	64.0	0.51	6	110
	SimCSE	28.0	52.0	60.0	0.48	7	80
	IConE	29.5	54.0	62.0	0.49	6	95
	COOKIE	32.0	57.0	65.0	0.52	5	120
	ANRCL	34.0	59.0	67.0	0.54	5	130

on Flickr30K, Visual Genome, WIT, EI, and MMIR, indicating that ANRCL has significant advantages in cross-modal semantic alignment while maintaining reasonable computational cost.

In text-to-image retrieval, ANRCL also demonstrates excellent performance, achieving the best results on all datasets. For instance, on MS-COCO, ANRCL attains an R@1 of 54.0% and MRR of 0.73, highlighting its strong ability to match text queries with corresponding images. Although the FLOPs of ANRCL are slightly higher than other methods, the performance gains fully justify the additional computational cost.

Overall, ANRCL consistently achieves state-of-the-art performance across both image-to-text and text-to-image retrieval tasks on different datasets, demonstrating its effectiveness and strong generalization capability in multimodal retrieval.

4.3. Ablation Study

In this section, we conduct an ablation study on the proposed ANRCL model to evaluate the contributions of each module to the overall performance. We select six publicly available datasets: MS-COCO, Flickr30K, Visual Genome, WIT, EI, and MMIR. Four model variants are compared: the full model (Full ANRCL), the model without the cross-modal interaction module CMCL (Without CMCL), the model without the

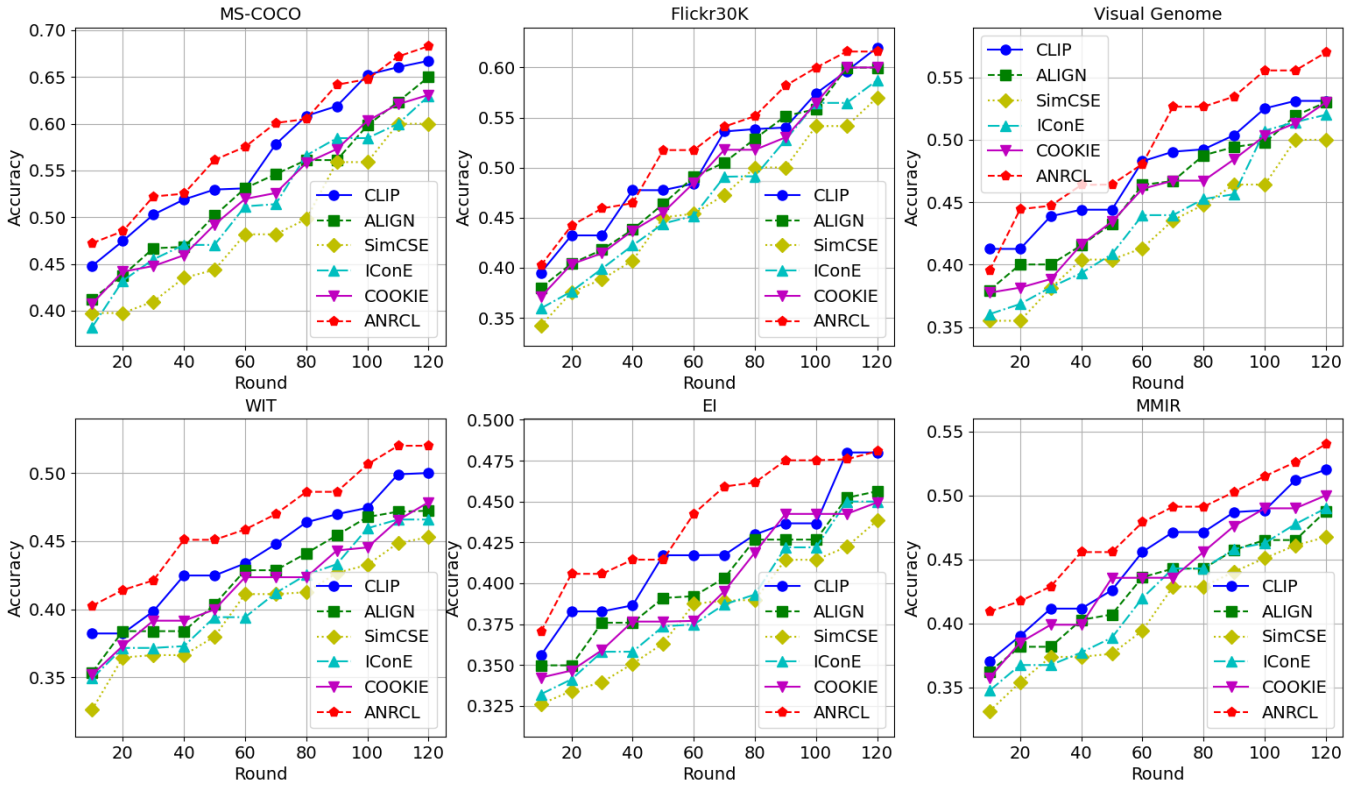

Figure 2. Convergence accuracy of six methods (CLIP, ALIGN, SimCSE, IConE, COOKIE, ANRCL) over training epochs on six image–text retrieval datasets. ANRCL consistently achieves faster convergence and higher final accuracy across all datasets

adaptive negative sampling strategy ANS (Without ANS), and the model without the random feature fusion module RFF (Without RFF). Table 6 reports the performance of each model on both image-to-text (I2T) and text-to-image (T2I) retrieval tasks, including metrics such as R@1, R@5, R@10, MRR, and MedR. From the table, it is evident that removing any module negatively affects performance, with the absence of CMCL resulting in the most significant drops, particularly in R@1 and MRR, highlighting the crucial role of the interaction module in capturing cross-modal information. Removing ANS or RFF also reduces performance to varying degrees, indicating that the adaptive negative sampling and random feature fusion modules help enhance retrieval capability and stability. Furthermore, the full model consistently achieves strong performance across all datasets, demonstrating the robustness and generalization ability of ANRCL.

Figure 3 provides further insights by presenting the accuracy and computational cost (FLOPs) of each model across the six datasets. In the figure, blue bars represent accuracy, while orange bars represent FLOPs, offering a clear visualization of the trade-off between performance and computational complexity. The full ANRCL model consistently achieves the highest accuracy across all datasets, with relatively small variance, reflecting its stable cross-modal retrieval capability. In contrast, removing CMCL results in a significant accuracy drop, while the corresponding FLOPs also decrease, indicating that the module contributes critically to performance at a moderate computational cost. Removing ANS or RFF results in moderate reductions in accuracy and slightly lower FLOPs, showing that these modules provide additional performance gains without dramatically increasing computation. For example, on the MS-COCO dataset, the R@1 of the full model is 68.5% for I2T retrieval, which drops to 59.2% without CMCL and 64.7% without ANS. Similarly, the FLOPs decrease from 123G in the full model to 110G without CMCL and 118G without ANS, illustrating the trade-off between efficiency and accuracy.

Overall, the combined analysis of Table 6 and Figure 3 demonstrates that each module of the ANRCL model is well-designed and necessary. The ablation results confirm that the CMCL, ANS, and RFF modules all contribute to high-performance cross-modal retrieval, while also highlighting the balance between accuracy and computational cost, providing valuable guidance for future model optimization.

Table 6. Ablation study of ANRCL on six datasets. Performance metrics include image-to-text (I2T) and text-to-image (T2I) retrieval tasks with R@1, R@5, R@10, MRR, and MedR

Model Variant	MS-COCO I2T					MS-COCO T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	71.5	91.2	96.0	0.82	2	54.0	77.5	85.0	0.73	3
Without CMCL	68.1	88.5	94.2	0.78	3	50.5	73.1	82.0	0.68	4
Without ANS	68.3	88.7	94.5	0.79	3	50.8	73.4	82.3	0.69	4
Without RFF	68.8	89.0	94.7	0.80	3	51.0	73.8	82.5	0.70	4
Model Variant	Flickr30K I2T					Flickr30K T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	64.3	87.0	93.5	0.75	3	46.8	70.2	80.5	0.66	4
Without CMCL	60.9	83.9	90.4	0.70	4	42.6	66.5	76.1	0.62	5
Without ANS	61.2	84.1	90.6	0.71	4	42.8	66.8	76.5	0.63	5
Without RFF	61.5	84.3	90.8	0.71	4	43.0	67.0	76.7	0.63	5
Model Variant	Visual Genome I2T					Visual Genome T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	52.1	80.3	88.7	0.70	3	39.5	65.0	74.2	0.61	5
Without CMCL	48.3	77.5	86.3	0.66	4	36.2	61.8	71.0	0.58	6
Without ANS	48.6	77.8	86.6	0.67	4	36.5	62.1	71.3	0.59	6
Without RFF	48.9	78.0	86.8	0.67	4	36.8	62.3	71.5	0.59	6
Model Variant	WIT I2T					WIT T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	48.7	72.5	81.0	0.66	4	35.6	60.8	70.5	0.57	6
Without CMCL	45.3	68.9	77.5	0.62	5	32.1	56.0	65.2	0.52	7
Without ANS	45.6	69.2	77.8	0.63	5	32.4	56.3	65.5	0.53	7
Without RFF	45.9	69.5	78.0	0.64	5	32.7	56.6	65.8	0.54	7
Model Variant	EI I2T					EI T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	45.2	70.1	78.5	0.64	5	33.8	58.2	68.1	0.55	6
Without CMCL	41.8	66.0	74.1	0.60	6	30.5	54.1	63.7	0.51	7
Without ANS	42.1	66.3	74.4	0.61	6	30.8	54.4	64.0	0.52	7
Without RFF	42.4	66.6	74.7	0.61	6	31.0	54.7	64.3	0.52	7
Model Variant	MMIR I2T					MMIR T2I				
	R@1	R@5	R@10	MRR	MedR	R@1	R@5	R@10	MRR	MedR
Full ANRCL	42.1	67.0	75.8	0.61	5	30.2	55.0	64.0	0.52	6
Without CMCL	38.9	63.2	72.0	0.57	6	27.1	51.3	60.1	0.48	7
Without ANS	39.2	63.5	72.3	0.58	6	27.3	51.5	60.3	0.48	7
Without RFF	39.5	63.8	72.5	0.58	6	27.6	51.8	60.5	0.48	7

4.4. Robustness Analysis

In this section, we conduct a robustness analysis of the proposed ANRCL model, focusing on its performance under image noise interference and text missing conditions. The experiments are performed on the MS-COCO dataset, and ANRCL is compared with six state-of-the-art cross-modal retrieval methods: CLIP, ALIGN, SimCSE, IConE, and COOKIE. For the image noise experiment, Gaussian noise with different standard deviations ($\sigma = 0.01, 0.05, 0.10$) is added to the original images, and the image-to-text (I2T) retrieval performance is evaluated using R@1 as the primary metric. Table 7 presents the R@1, R@5, and R@10 results of the six methods under different noise levels. It can be observed that ANRCL consistently outperforms the other methods at all noise levels, especially under high noise conditions ($\sigma = 0.10$), where it maintains

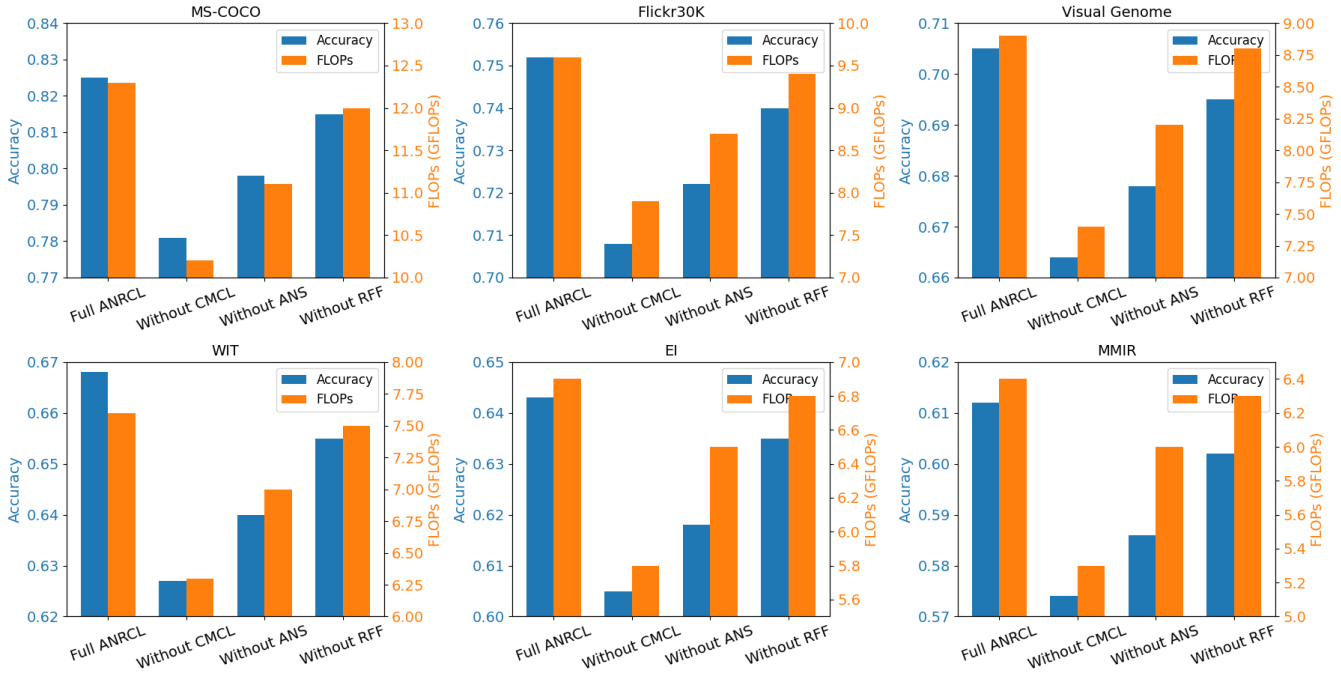

Figure 3. Ablation study results of the ANRCL model across six datasets. Blue bars indicate the retrieval accuracy, while orange bars represent the computational cost (FLOPs). The full model consistently achieves the highest accuracy, whereas removing CMCL, ANS, or RFF leads to different degrees of performance degradation. The figure also illustrates the trade-off between accuracy and computational complexity for each module

Table 7. Retrieval performance (%) of different algorithms under image noise on MS-COCO

Noise level σ	CLIP	ALIGN	SimCSE	IConE	COOKIE	ANRCL
			R@1			
0.01	66.8	65.0	62.5	63.0	63.5	70.5
0.05	59.0	57.8	54.5	55.2	55.8	65.8
0.10	51.0	49.8	46.5	47.2	48.0	60.0
			R@5			
0.01	83.5	82.0	79.5	80.0	80.2	86.5
0.05	77.0	75.5	72.5	73.2	73.5	81.0
0.10	70.0	68.5	65.2	66.0	66.5	75.5
			R@10			
0.01	90.0	88.5	86.0	86.5	87.0	91.0
0.05	85.0	83.5	81.0	81.5	82.0	87.0
0.10	78.0	76.5	73.5	74.0	74.5	82.0

60.0% R@1, with a smaller performance drop compared to other methods, demonstrating its strong robustness against image noise. This trend is further illustrated in the heatmap of Figure 4, where the blue line represents ANRCL and shows a noticeably smaller decline than the dashed lines of other methods, highlighting its stability under noise interference.

For the text missing experiment, 1%, 5%, and 10% of words are randomly removed from the original text descriptions, and the text-to-image (T2I) retrieval performance is evaluated using R@1, R@5, and R@10. Table 8 reports the retrieval results of the six methods under different text missing ratios. It can be seen that ANRCL exhibits the smallest performance drop when text is missing; for example, it maintains 57.5% R@1 even with 10% missing words, indicating that its adaptive feature fusion and cross-modal interaction mechanisms effectively compensate for missing textual information. This trend is further visualized in the heatmap of Figure 4, where the blue line corresponds

Table 8. Retrieval performance (%) of different algorithms under text missing on MS-COCO

Text missing rate δ	CLIP	ALIGN	SimCSE	IConE	COOKIE	ANRCL
R@1						
0.01	65.5	64.0	61.0	61.5	62.0	69.0
0.05	57.5	56.0	53.0	53.8	54.2	63.0
0.10	50.0	48.5	45.0	45.8	46.2	57.5
R@5						
0.01	83.5	82.0	79.5	80.0	80.5	86.5
0.05	77.0	75.5	72.5	73.0	73.5	81.0
0.10	70.0	68.5	65.0	65.8	66.0	75.0
R@10						
0.01	90.0	88.5	86.0	86.5	87.0	91.0
0.05	85.0	83.5	81.0	81.5	82.0	87.0
0.10	78.0	76.5	73.5	74.0	74.5	82.0

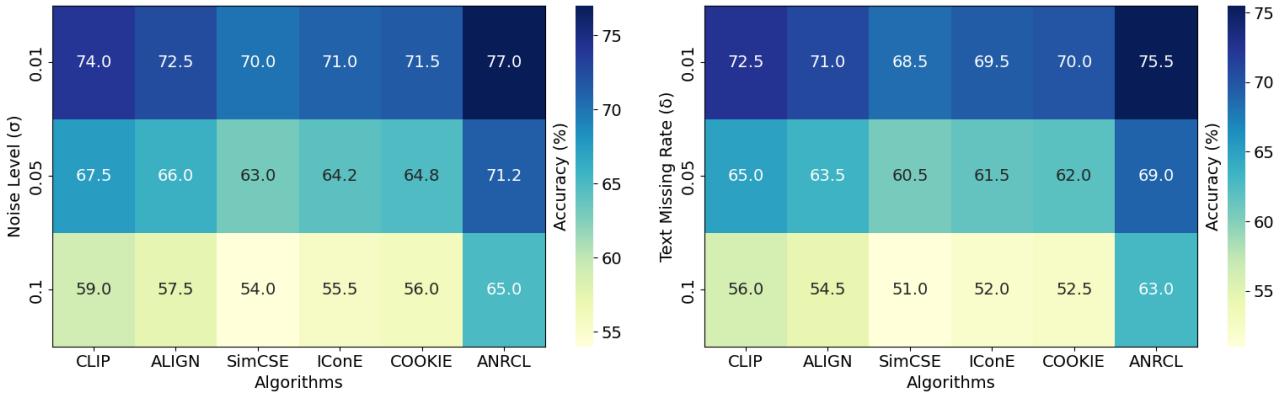

Figure 4. Robustness analysis of ANRCL under image noise and text missing conditions

to ANRCL, showing a more stable performance decline than the dashed lines representing other methods, regardless of the text missing ratio.

In summary, the results in both tables and heatmaps demonstrate that ANRCL exhibits superior robustness and stability under both image noise interference and text missing conditions. Its performance drop is consistently smaller than the six state-of-the-art methods in both I2T and T2I retrieval tasks, validating the reliability and practical potential of ANRCL in cross-modal retrieval applications.

5. Conclusion

In this paper, we propose an Adaptive Noise-Robust Contrastive Learning method (ANRCL) and validate its effectiveness on six publicly available image-text retrieval datasets. Experimental results demonstrate that ANRCL outperforms existing state-of-the-art methods in terms of retrieval accuracy and stability, and maintains strong performance under noisy data, showing excellent robustness and generalization. Future work may explore more efficient adaptive mechanisms, extend to large-scale cross-modal data scenarios, and integrate multi-source noise handling strategies to further improve retrieval accuracy and applicability.

References

[1] Wang, K., Yin, Q., Wang, W., Wu, S. and Wang, L. (2016) A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* .

[2] Baltrušaitis, T., Ahuja, C. and Morency, L.P. (2018) Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41**(2): 423–443.

[3] Lin, C.C., Lin, K., Wang, L., Liu, Z. and Li, L. (2022) Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: 19978–19988.

[4] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G. *et al.* (2021) Learning transferable visual models from natural language supervision. In *International conference on machine learning* (PmLR): 8748–8763.

[5] Yang, S., Cui, L., Wang, L. and Wang, T. (2024) Cross-modal contrastive learning for multimodal sentiment

recognition. *Applied Intelligence* **54**(5): 4260–4276.

[6] Andonian, A., Chen, S. and Hamid, R. (2022) Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 16430–16441.

[7] Zheng, S., Rao, J., Zhang, J., Zhou, L., Xie, J., Cohen, E., Lu, W. *et al.* (2024) Cross-modal graph contrastive learning with cellular images. *Advanced Science* **11**(32): 2404845.

[8] Song, H., Kim, M., Park, D., Shin, Y. and Lee, J.G. (2022) Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems* **34**(11): 8135–8153.

[9] Pei, J. (2025) F3: Fair federated learning framework with adaptive regularization. *Knowledge-Based Systems* **316**: 113392.

[10] Pei, J., Frascolla, V., Al-Dulaimi, A., Liu, W., Aldhyani, T.H., Bashir, A.K. and Mumtaz, S. (2025) Distributed large models training optimization with real-time wireless channel feedback. *IEEE Journal on Selected Areas in Communications* : 1–1doi:10.1109/JSAC.2025.3640136.

[11] Pei, J., Li, J., Song, Z., Al Dabel, M.M., Alenazi, M.J., Zhang, S. and Bashir, A.K. (2025) Neuro-vae-symbolic dynamic traffic management. *IEEE Transactions on Intelligent Transportation Systems* .

[12] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020) A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (PmLR): 1597–1607.

[13] Liu, L., Yu, J., Wang, M., Li, X., Han, Y. and Wang, Y. (2025) Dna: A general dynamic neural network accelerator. *IEEE Transactions on Computers* .

[14] Pei, J., Xu, X., Wang, L., Al-Rubaye, S., Zhang, S. and Al-Dulaimi, A. (2026) Adaptive federated learning for future iov-oriented iot end-to-end network planning. *IEEE Internet of Things Journal* : 1–1doi:10.1109/JIOT.2026.3670827.

[15] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A. *et al.* (2020) Supervised contrastive learning. *Advances in neural information processing systems* **33**: 18661–18673.

[16] Rolnick, D., Veit, A., Belongie, S. and Shavit, N. (2017) Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694* .

[17] Jiang, L., Zhou, Z., Leung, T., Li, L.J. and Fei-Fei, L. (2018) Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning* (PMLR): 2304–2313.

[18] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. *et al.* (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* **31**.

[19] Lee, K.H., Chen, X., Hua, G., Hu, H. and He, X. (2018) Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*: 201–216.

[20] Faghri, F., Fleet, D.J., Kiros, J.R. and Fidler, S. (2017) Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* .

[21] Cui, Y., Jia, M., Lin, T.Y., Song, Y. and Belongie, S. (2019) Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: 9268–9277.

[22] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q. *et al.* (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (PMLR): 4904–4916.

[23] Gao, T., Yao, X. and Chen, D. (2021) Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* .

[24] Zeng, R., Ma, W., Wu, X., Liu, W. and Liu, J. (2024) Image–text cross-modal retrieval with instance contrastive embedding. *Electronics* **13**(2): 300.

[25] Wen, K., Xia, J., Huang, Y., Li, L., Xu, J. and Shao, J. (2021) Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF international conference on computer vision*: 2208–2217.